On the Minimal Degree Bias in Generalization on the Unseen for non-Boolean Functions

Denys Pushkin[1], Raphaël Berthier[2], Emmanuel Abbe[1 3]

[1]EPFL, [2]Inria Sorbonne Université, [3]Apple

Generalization on the Unseen

Consider the target function $f(x_1, \ldots, x_d) = x_1 \cdot x_2$ on the training domain $\{x \in \Omega \mid (x_1 - 1)(x_2 - 1) = 0\}$. Where will the model converge on the unseen part of the domain? As shown by *[Abbe et al., 2023]*, when considering the boolean domain $\Omega = \{\pm 1\}^d$ and sparse regime ($d \to \infty$), a set of models including the Random Feature model and Transformer converge to the minimum-degree interpolator (MDI), which is given in this case by $x_1 + x_2 - 1$.

Main Question

Does the min-degree bias extend beyond the boolean domains?

Random Feature Model

Random Features (RF) model: $f_{\text{RF}}(a; x) = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_i \phi_{w_i, b_i}(x)$, $x \in \mathbb{R}^d$, where $\phi_{w,b}(x) = \sigma(\langle w, x\rangle + b)$ are the random features. Here, only parameter $a \in \mathbb{R}^d$ is trainable, while parameters $\{w_i\}_{i=1}^N$ and $\{b_i\}_{i=1}^N$ are sampled randomly and then fixed during the training.

Sparse regime: $w_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$, $b \sim \mathcal{N}(0, \frac{1}{d})$, and $d \to \infty$.

Small feature regime: $w_i \sim \mathcal{N}(0, \varepsilon I_d)$, $b \sim \mathcal{N}(0, \varepsilon)$, $\varepsilon \to 0$.

RF Model in Sparse Regime Breaks the MDI Bias in Real-Valued Domain

Table: Training the Random Feature model on $f(x) = 1$, $x \in \mathbb{R}^d$ with GOTU constraint $x_1 = 1$ in sparse regime. Here, $d = 15$ and $N = 1024$. In the second and the third column, you can see the monomial coefficient learnt by model.

Activation	1	x_1
$(1+x)^2$	0.624 ± 0.017	0.374 ± 0.017
ReLU	0.564 ± 0.009	0.431 ± 0.011
Shifted ReLU	0.782 ± 0.009	0.217 ± 0.011
Sigmoid	0.992 ± 0.003	0.007 ± 0.002
Softplus	0.789 ± 0.010	0.208 ± 0.012

RF Model in Small Feature Regime Preserves MDI Bias in Real-Valued Domain

Main Theorem (Informal)

As the number of random features $N \to \infty$ before the random features scale $\varepsilon \to 0$, the Random Feature model with polynomial activation converges to some interpolator of minimum degree in small feature regime.

Table: Training the Random Feature model on $f(x) = 1$, $x \in \mathbb{R}^d$ with GOTU constraint $x_1 = 1$ in small feature regime with $\varepsilon = (0.03)^2$. Here, $d = 15$ and $N = 1024$.

Activation	1	x_1
$(1+x)^2$	0.997 ± 0.002	0.001 ± 0.003
ReLU	0.564 ± 0.009	0.430 ± 0.010
Shifted ReLU	1.000 ± 0.000	-0.001 ± 0.003
Sigmoid	1.000 ± 0.000	-0.001 ± 0.003
Softplus	1.000 ± 0.001	-0.001 ± 0.003

Motivation for Small Feature Regime

Consider the setting of multi-index model where we learn the target of the form

$$f(x) = \varphi(U^\top x)$$

where $f : \mathbb{R}^d \to \mathbb{R}$, $\varphi : \mathbb{R}^k \to \mathbb{R}$, $U \in \mathbb{R}^{d \times k} :\ U^\top U = I_k$, k is fixed dimension and $d \gg 1$ is large dimension.

Define the loss function as $\mathcal{L}(a) = \frac{1}{2}\mathbb{E}_x \left[\left(f(x) - f_{\text{RF}}(a; x)\right)^2 \right]$.

Proposition

$$\mathcal{L}(a) = \frac{1}{2}\mathbb{E}_z \left[\left(\varphi(z) - \sum_{i=1}^N a_i \overline{\sigma}_i(\langle U^\top w_i, z\rangle + c_i) \right)^2 \right] + \frac{1}{2} a^\top \Lambda a \,,$$

where $z = U^\top x$ and $\Lambda \succeq 0$.

Intuition: high-dimensional regression problem in $x \in \mathbb{R}^d$ reduces to lower-dimensional in $z \in \mathbb{R}^k$ with an additional regularizer term $\frac{1}{2} a^\top \Lambda a$. Moreover, if $w_i \sim \mathcal{N}(0, \frac{1}{d} I_d)$, then $U^\top w_i \sim \mathcal{N}(0, \frac{1}{d} I_k)$, and the latter corresponds to the small feature regime.

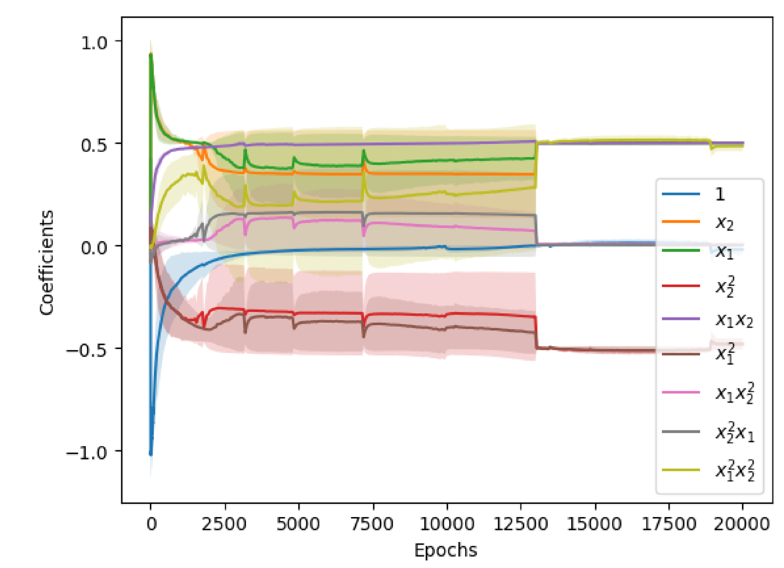
MDI for Data Embedded in Roots of Unity

Consider learning the target function $f : \mathbb{U}_n^d \to \mathbb{C}$, where $\mathbb{U}_n$ denotes n-roots of unity, using the complex Random Feature model, for which $a_i \in \mathbb{C}$, $w_i \sim \mathcal{CN}(0, \frac{1}{d} I_d)$, $b_i \sim \mathcal{CN}(0, \frac{1}{d} I_d)$, where $\mathcal{CN}$ is complex standard normal distribution.

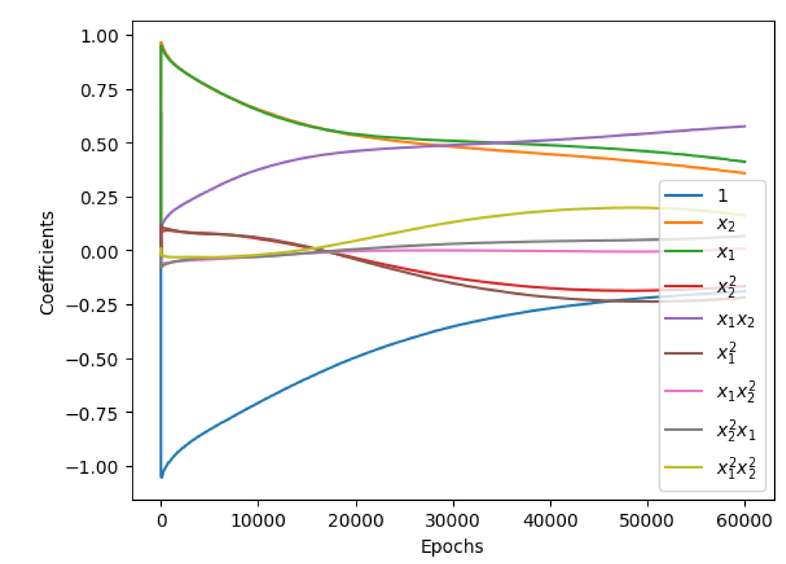
Theorem (Informal)

As the number of random features $N \to \infty$ before the dimension $d \to \infty$, the complex Random Feature model converges to the interpolator of minimum degree.

What about Transformer?



lr $= 10^{-4}$



lr $= 10^{-5}$

Training Transformer on $f(x) = x_1 x_2$, $x \in \{-1, 0, 1\}^d$ with GOTU constraint $(x_1 - 1)(x_2 - 1) = 0$ in dimension $d = 15$ using AdamW optimizer. The MDI is given by $x_1 + x_2 - 1$, but the Transformer (with $lr = 10^{-4}$) converges close to $f_{int}(x) = \frac{1}{2}(x_1 + x_2 - x_1^2 + x_1x_2 - x_2^2 + x_1^2x_2^2)$.

Future Work

- What if not min-degree bias governs the generalization of the Random Feature and Transformer models on the real-valued domains?

Related Work

- *[Abbe et al., 2023]*: Generalization on the Unseen, Logic Reasoning and Degree Curriculum.